BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 7: e112436
doi: 10.3897/biss.7.112436

OPEN ACCESS

Conference Abstract

Unearthing the Past for a Sustainable Future: Extracting and transforming data in the Biodiversity Heritage Library for climate action

JJ Dearborn‡, Mike Lichtenberg‡, Joel Richard‡, Joseph deVeer§, Michael Trizna|, Katie Mika¶

‡ Smithsonian Libraries and Archives, Biodiversity Heritage Library, Washington, D.C., United States of America
§ Harvard University, Museum of Comparative Zoology, Ernst Mayr Library, Cambridge, MA, United States of America
| Smithsonian Institution, Office of the Chief Information Officer, Data Science Lab, Washington, United States of America
¶ Harvard Library & Institute for Quantitative Social Science, Cambridge, Massachusetts, United States of America

Corresponding author: JJ Dearborn (dearbornjj@si.edu)

Received: 08 Sep 2023 | Published: 11 Sep 2023

Citation: Dearborn J, Lichtenberg M, Richard J, deVeer J, Trizna M, Mika K (2023) Unearthing the Past for a Sustainable Future: Extracting and transforming data in the Biodiversity Heritage Library for climate action. Biodiversity Information Science and Standards 7: e112436. https://doi.org/10.3897/biss.7.112436

Abstract

As the urgency to address the climate crisis intensifies, the availability of accurate and comprehensive biodiversity data has become crucial for informing climate change studies, tracking key environmental indicators, and building global biodiversity monitoring platforms. The Biodiversity Heritage Library (BHL) plays a vital role in the core biodiversity infrastructure, housing over 60 million pages of digitized literature about life on Earth. Recognizing the value of over 500 years of data in BHL, a global network of BHL staff is working to establish a scalable data pipeline to provide actionable occurrence data from BHL's vast and diverse collections. However, transforming textual content into FAIR (findable, accessible, interoperable, reusable) data poses challenges due to missing descriptive metadata and error-ridden unstructured outputs from commercial text engines. (Fig. 1)

Despite the wealth of knowledge in BHL now available to global audiences, the underutilization of biodiversity and climate data contained in BHL's textual corpus hinders scientific research, hampers informed decision-making for conservation efforts, and limits our understanding of biodiversity patterns crucial for addressing the climate crisis. By

This is an open access article distributed under the terms of the CC0 Public Domain Dedication.

leveraging recent advancements in text recognition engines, along with cutting-edge AI (Artificial Intelligence) models like OpenAI's CLIP (Contrastive Language-Image Pre-Training) and nascent features in transcription platforms, BHL staff are beginning to process vast amounts of textual and image data and transform centuries worth of data from BHL collections into computationally usable formats. Recent technological breakthroughs now offer a transformative opportunity to empower the global biodiversity community with prescient insights from our shared past and facilitate the integration of historical knowledge into climate action initiatives.

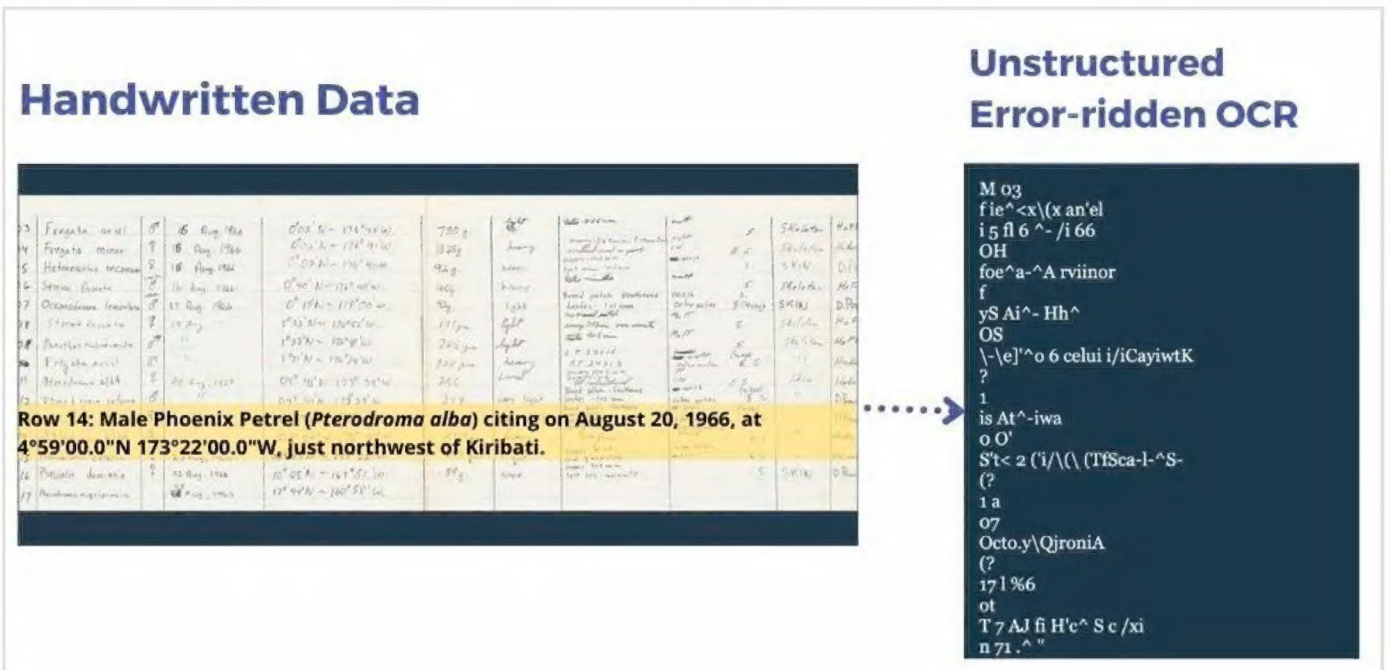

Figure 1.

Sample of handwritten observation data with corresponding unstructured, uncorrected OCR (optical character recognition) text. From National Museum of Natural History, Pacific Ocean Biological Survey Program, At-sea, 1963-1966, 1968, part 3: July - August 1966. Image credit: Dearborn, 2023 | Creative Commons Attribution 4.0 license (CC-BY).

To bridge gaps in the historical record and unlock the potential of the Biodiversity Heritage Library (BHL), a multi-pronged effort utilizing innovative cross-disciplinary approaches is being piloted. These technical approaches were selected for their efficiency and ability to generate rapid results that could be applied across the diverse range of materials in BHL. (Fig. 2)

Piloting a data pipeline that is scalable to 60 million pages requires considerable investigation, experimentation, and resources but will have an appreciable impact on global conservation efforts by informing and establishing historic baselines deeper into time. This presentation will focus on the identification, extraction, and transformation of OCR into structured data outputs in BHL. Approaches include:

- Upgrading legacy OCR text using Tesseract OCR engine to improve data quality by 20% and openly publish 40 GBs of textual data as FAIR data;
- Evaluating handwritten text recognition (HTR) engines (Microsoft Azure Computer Vision, Google Cloud Vision API (Application Programming Interface), and Amazon

Textract) to improve scientific name-finding in BHL's handwritten archival materials using algorithms developed by Global Names Architecture;

- Extracting data from collecting events using HTR coordinate outputs with Python library Pandas DataFrame to create structured data;
- Classifying BHL page-level images with OpenAI's CLIP, a neural network model to accurately identify the handwritten sub-corpus of primary source materials in BHL;
- Running an A/B test to evaluate the efficiency and accuracy of human-keyed transcription data extraction to provide high-quality, human-vetted datasets that can be deposited with data aggregators.

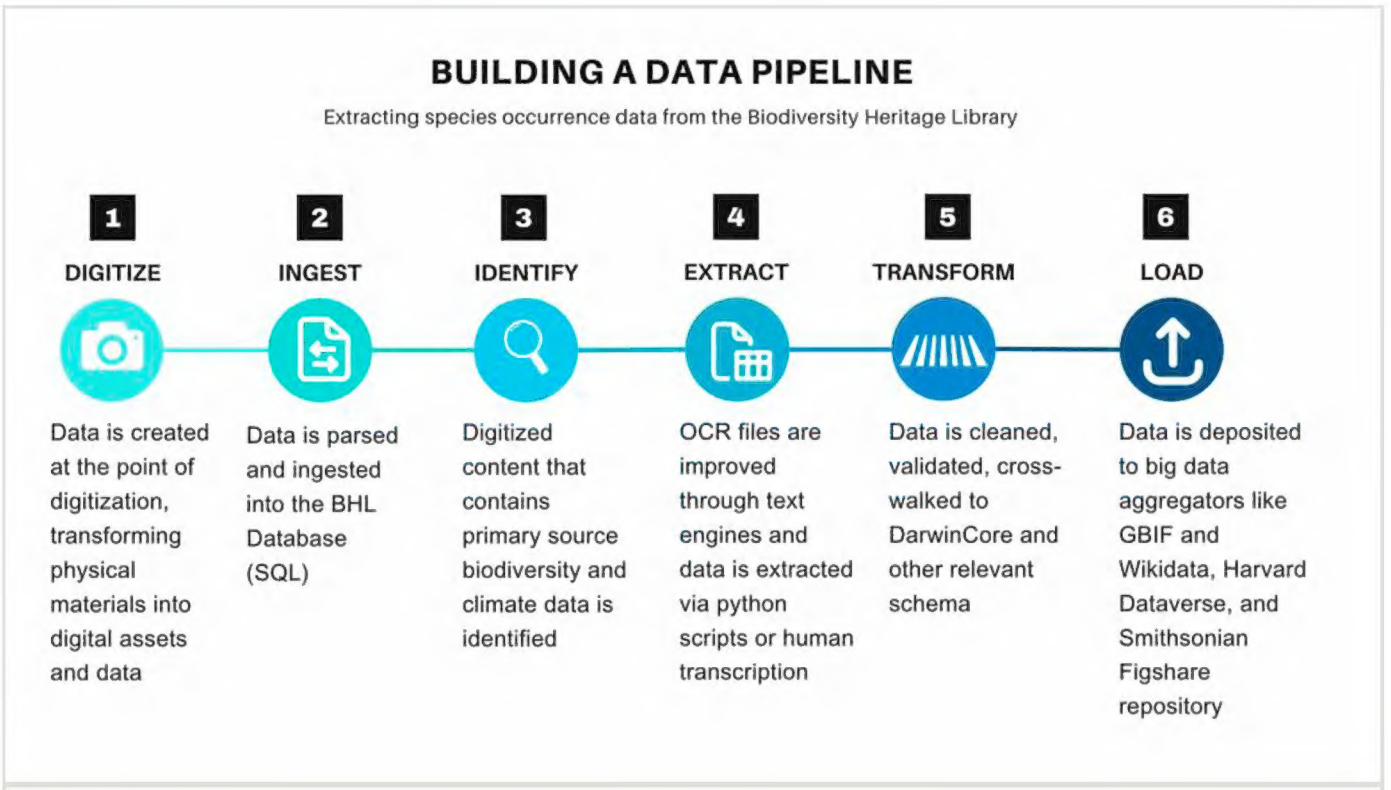
BUILDING A DATA PIPELINE
Extracting species occurrence data from the Biodiversity Heritage Library
1
DIGITIZE
Data is created at the point of digitization, transforming physical materials into digital assets and data
2
INGEST
Data is parsed and ingested into the BHL Database (SQL)
3
IDENTIFY
Digitized content that contains primary source biodiversity and climate data is identified
4
EXTRACT
OCR files are improved through text engines and data is extracted via python scripts or human transcription
5
TRANSFORM
Data is cleaned, validated, cross-walked to DarwinCore and other relevant schema
6
LOAD
Data is deposited to big data aggregators like GBIF and Wikidata, Harvard Dataverse, and Smithsonian Figshare repository

Figure 2.

Six steps to building a data pipeline for species occurrence data from BHL to data aggregators. Image credit: Dearborn, 2023 | Creative Commons Attribution 4.0 license (CC-BY).

The ongoing development of a scalable data pipeline of BHL's relevant biodiversity and climate-related datasets requires sustained support and partnership with the biodiversity community. Initial results demonstrate that liberating data from archival and handwritten field notes is arduous but feasible. Extending these methodologies to the broader scientific literature presents new research opportunities. Extracting and normalizing data from unstructured textual sources can significantly advance biodiversity research and inform environmental policy. The Biodiversity Heritage Library staff are committed to building multiple scalable data pipelines with the ultimate goal of erecting a global biodiversity knowledge graph, rich in interconnected data and semantic meaning, enabling informed decisions for the preservation and sustainable management of Earth's biodiversity.

Keywords

Global Names Architecture (GNA), global biodiversity infrastructure, scalable data pipelines, historic literature, species occurrence data, climate change, handwritten text recognition, image classification, crowdsourced transcription, structured data, global biodiversity community, data extraction, machine learning, artificial intelligence

Presenting author

JJ Dearborn

Presented at

TDWG 2023

Acknowledgements

Many thanks to additional collaborators from BHL's partner network for making this research possible:

BHL Transcription Upload Tool Working Group (BHL-TUTWG). Members listed in alphabetical order: Riccardo Ferrante (Smithsonian Libraries and Archives), Kelly Hall (Auckland War Memorial Museum, Tamaki Paenga Hira or Auckland Museum), David Iggulden (Kew Gardens Library and Archives), Susan Lynch (BHL), Diane Rielinger (Harvard Botany Libraries), Gretchen Rings (Field Museum, Chicago), Rebekah Kim (California Academy of Sciences), Judy Warnement (BHL).

Global Names Architecture (GNA): Dr. Dmitry Mozzherin, Scientific Informatics Leader at Marine Biological Laboratory and Dr. Geoff Ower, Research Programmer at Illinois Natural History Survey.

Hosting institution

Major support and hosting is provided by Smithsonian Libraries and Archives.

Conflicts of interest

The authors have declared that no competing interests exist.